DOCUMENT RESUME

ED 390 900 TM 024 213

AUTHOR	Bennett, Randy Elliot; And Others
TITLE	Clusters as the Unit of Analysis in Differential Item Functioning.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-88-70
PUB DATE	Dec 88
NOTE	54p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Black Students; Braille; *Cluster Analysis; *Difficulty Level; High Schools; *High School Students; Hypothesis Testing; Item Analysis; *Item Bias; Mathematics Tests; *Test Items; Test Theory; Visual Impairments
IDENTIFIERS	Scholastic Aptitude Test

ABSTRACT

This study developed, applied, and evaluated a theory-based method of detecting the underlying causes of differential difficulty. The method was applied to two subgroups taking the Scholastic Aptitude Test-Mathematics (SAT-M), 261 visually impaired students taking Braille forms of the test and 1,985 black students at 3 test administrations. It involved: (1) reviewing the literature to identify possible causes of differential item functioning; (2) forming item categories based on those factors; (3) identifying categories that functioned differentially; (4) assessing the functioning of the items composing deviant categories; and (5) relating item and category functioning. Results were compared to a traditional item-level analysis. In both subgroups, the cluster and traditional methods agreed on the overall extent of differential functioning (substantial in the first group, virtually absent in the second). The method would seem to be applied most productively when a small number of hypotheses can be derived from a reasonably strong research base, overlap among cluster structures can be avoided, and results can be supplemented with experimental studies of protocol analysis. (Contains 27 references, 11 tables, and an appendix of category definitions.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

CLUSTERS AS THE UNIT OF ANALYSIS IN DIFFERENTIAL ITEM FUNCTIONING

Randy Elliot Bennett
Donald A. Rock
Inge Novatkoski

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
December 1988

Clusters as the Unit of Analysis in Differential Item Functioning

Randy Elliot Bennett

Donald A. Rock

and

Inge Novatkoski

Copyright © 1988. Educational Testing Service. All rights reserved.

Acknowledgements

Gratitude is expressed to Richard Fiddler, Dolores Godzeiba and Jim Braswell for their input on potential sources of differential difficulty for blind students taking brailled mathematics tests, to David Wright for his assistance in locating SAT score files and related information, and to Jane Kupin and Margaret Thorpe for their help in classifying items. Peter Pashley and Charlie Lewis provided helpful comments on an earlier draft of this report. Finally, thanks are owed to Warren Willingham for his help and encouragement, and to the Program Research Planning Council for its support.

Abstract

This study developed, applied, and evaluated a theory-based method of detecting the underlying causes of differential difficulty. The method was intended to improve on traditional approaches that too often produce uninterpretable results. Key elements of the method were the analysis of item clusters and the incorporation of theoretical predictions about cluster performance. The method was applied in two subgroups taking SAT-M and involved (1) reviewing literature syntheses to identify factors that might cause differential item functioning, (2) forming item categories based on those factors, (3) identifying categories that functioned differentially, (4) assessing the functioning of the items composing deviant categories, and (5) relating item and category functioning. Results were compared to a traditional item-level analysis. In both subgroups, the cluster and traditional methods agreed on the overall extent of differential functioning (substantial in the first group, virtually absent in the second). Additionally, the pattern of differential functioning detected was interpretable. At the same time, several important limitations were apparent. The method would seem to be applied most productively when a small number of hypotheses can be derived from a reasonably strong research base, overlap among cluster structures can be avoided, and results can be supplemented with experimental studies or protocol analyses.

Clusters as the Unit of Analysis in Differential Item Functioning

Most traditional approaches to differential item functioning are built on the evaluation of relatively unreliable, individual items (individual items are unreliable in that they are generally poor indicators of the content universe or construct they are intended to measure). When differential functioning is being studied for several groups simultaneously (e.g., Black examinees, women, Hispanic candidates, handicapped people), or when subgroup sample sizes are small, this limitation can have particularly severe consequences. Because individual items are relatively unreliable, numerous test questions will show statistically significant evidence of differential functioning by chance alone. Interpreting on a post hoc basis the resulting mix of false and true-positive items has proved very difficult, resulting in little success locating the factors underlying differential functioning across test questions.

To increase the chances of identifying underlying factors, investigators have implemented both experimental, theory-based methods and approaches based on the analysis of item clusters (e.g., Scheuneman, 1985; Schmeiser, 1981; Wild, 1987a). Building upon these conceptual advances, we propose an a priori, theory-based method built upon item clusters (see Bennett, Rock, & Kaplan, 1987, for an initial version of the methodology). Clusters are suggested as the unit of analysis because they are more reliable than individual items (more

indicators of the universe are presented). A theory-based approach is advocated because it forces hypotheses about differential functioning to be stated in advance of the data analysis. In combination, the use of clusters and theory should reduce the frequency of false positives, make more systematic the search for underlying causes, and provide better information for program policy decisions concerning the modification or possible elimination of broad item classes found to operate differentially for one or another group. The purpose of this study was to develop, try out, and evaluate such a theory-based method using items from the Scholastic Aptitude Test (SAT).

Subjects

Subjects were members of two groups for whom differential difficulty currently is a concern. For the first group, visually impaired students taking the braille edition of the SAT, instances of differential difficulty have been found on the test's Mathematical section (Bennett, Rock, & Kaplan, 1987). These instances did not appear to be associated with items possessing any single, distinctive characteristic. Hence, a closer look at the performance of this group seems warranted.

Differential difficulty for visually impaired candidates is particularly hard to evaluate via traditional methods because so few of these examinees take the braille edition and because content, format, and administration (i.e., timing) effects may be confounded. On the first count, the group

poses an appropriate methodological challenge because of the increased stability associated with analyzing clusters instead of items, an increase that should be particularly valuable with small samples. The potential confounding, however, poses a serious problem--as it would for any traditional approach--and it may be that the causes of any significant observed effects can only be fully understood through experimental manipulations.

The second group is composed of Black students. Numerous studies have focused upon the functioning of SAT Verbal items with several investigations finding evidence of differential difficulty (e.g., Dorans, 1982; Kulick, 1984; Rogers & Kulick, 1986; Schmitt, Bleistein, & Scheuneman, 1987). Far less attention has been paid, however, to the performance of this group on the Mathematical section. That differential difficulty might occur on the Mathematical section is suggested by the results of research on other mathematics tests taken by high school and college-age populations (Scheuneman, 1978, 1985; Shepard, Camilli, & Williams, 1984).

Visually impaired subjects were drawn from a pool of students taking special, extended-time administrations of SAT forms WSA3, WSA5, and CSA5. The WSA forms were administered from March 1980 through June 1983 and the CSA form from October 1983 through September 1986. (A second CSA form, CSA7, was taken by too few visually impaired students to make analysis worthwhile.) Visually impaired students were eliminated from the pool if they requested a special test

edition other than braille (e.g., cassette) or requested the braille edition in conjunction with another edition (e.g., large type). (Those who requested the braille edition along with a regular print copy for use by a reader were retained.) Further, students were eliminated if they indicated on the Student Descriptive Questionnaire that English was not their best language. The resulting samples consisted of 91 students for WSA3, 96 for WSA5, and 74 for CSA5.

The performance of each of these handicapped samples was compared to a random sample of high school students who took the regular print versions of the same test forms under standard conditions. For WSA3, the reference group consisted of 1,110 students randomly drawn from a two-state administration in October 1974; for WSA5, 1,398 examinees were selected randomly from the equating bank for a national administration in December of that same year (equating banks are large random samples used for placing forms on the SAT score scale). The CSA5 sample also was drawn randomly from the equating bank for a national administration in October 1980; 5,507 students composed this sample. All three samples were drawn to conform to the proportions of seniors in the handicapped groups: approximately .73 for WSA3, .05 for WSA5, and .91 for CSA5. As for visually impaired examinees, students indicating that English was not their best language were eliminated from the reference samples.

Black subjects were randomly selected from those students taking standard administrations of three test forms: CSA5,

CSA7, and GSA2. The CSA5 national administration date is given above, while CSA7 was administered nationally in December 1980 and GSA2 in January 1984. Random samples of approximately 5,000 examinees were drawn from the appropriate equating banks. These samples were then separated into Black and White groups. Finally, the White sample was adjusted (by deleting examinees) to produce the same proportions of juniors and seniors as the Black group. The resulting samples were for Black examinees, 446, 834, and 705, and for White examinees, 4,405, 4,798, and 3,985, respectively. Proportions of seniors were approximately .90, .79, and .96 for CSA5, CSA7, and GSA2, respectively.

Tables 1 and 2 present background information on the study and reference groups. For the visually impaired samples, perhaps the most obvious characteristic is their extremely small size--even though data were pooled across the three years that each form remained in service. Second, this group is consistently older than the reference sample, suggesting that these students take longer to progress through school. Finally, their SAT-M scores are consistently and substantially lower than their nonhandicapped peers (though by different amounts) while their SAT-V scores are not.

Insert Tables 1 and 2 about here

For the Black examinee samples, several characteristics stand out. First, males are underrepresented relative to the

reference group. Second, as in many previous analyses, Mathematical and Verbal scores, and self-reported math grades, are noticeably lower than the reference group's values (for scores, the differences approximate 1 standard deviation while for grades, they range from .20 to .37 standard deviations). Finally, the number of years of math taken is somewhat greater for the White group on two of the three test forms. This last result is particularly noteworthy as many studies have documented lower Black enrollments especially in precollege math courses (e.g., Johnson, 1984; Jones, 1984; Jones, Burton, & Davenport, 1984; Matthews, Carpenter, Lindquist, & Silver, 1984). For students taking the SAT, a self-report of specific course type was not requested until the advent of the GSA forms. Review of these data confirm that Black students administered GSA2 take fewer years of advanced math courses (i.e., trigonometry, precalculus, calculus) and more years of "other" math courses than their White peers.

Method

This study involved two major steps. First, each set of test forms was analyzed using the cluster-based method. Second, the method was evaluated to determine its utility for studying differential functioning.

The Cluster-Based Method

The cluster-based method involved (1) reviewing literature syntheses to identify factors that might cause differential item functioning, (2) forming item categories based on these factors, (3) identifying categories that

functioned differentially, (4) assessing the functioning of the items composing errant categories (when such categories were found), and (5) relating item and category functioning (again, conditional on discovering deviant categories).

Identifying relevant subgroup factors. To identify subgroup factors relevant to SAT Mathematical item performance, several activities were undertaken. First, searches were conducted of the ERIC and Psychological Abstracts databases to identify for each study group syntheses of the literature on mathematics and cognitive processing. Second, existing differential item functioning studies were reviewed for indications of how subgroup characteristics might affect item performance. Finally, individuals knowledgeable about the characteristics of the subgroups were contacted for suggestions.

For both groups, these activities produced a limited amount of information. In particular, the literature searches uncovered surprisingly few relevant references. For visually impaired students, the available sources indicated three main factors that might produce differential functioning. The first factor related to the cognitive characteristics of visually impaired students. The major difference in cognitive characteristics between sighted students and those blind from birth relates, of course, to visual experience. Because of their more limited experience, blind students generally have less well-developed spatial abilities than their sighted peers. Differences in ability may manifest themselves on such

tasks as synthesizing shapes from their component parts (Warren, 1981), perceiving embedded figures (Witkin, Birnbaum, Lomonaco, Lehr, & Herman, 1968), using or understanding spatial language (Warren, 1981), or working with complex figures (e.g., figures in three dimensions; figures with shaded portions in which the dots used for shading may be misinterpreted as braille notation).

A second factor is that the availability of sight makes some operations easier. For example, sighted examinees frequently can gather visually some of the information needed to arrive at a solution; for instance, the sizes of "special" angles (i.e., 30°, 45°, 60°) can be estimated and compared. Further, sighted students can sometimes eliminate incorrect options through visual inspection or even tentatively identify the correct one. Finally, they can construct diagrams as an aid in solving certain types of items (e.g., Venn diagrams for logic problems).

The third factor was related to braille. Reading in braille has associated with it several pertinent effects. First, text takes more space to represent. One result is visually impaired students who use this medium are generally slower readers than nonhandicapped students. Roman numeral item formats (i.e, where the answer options refer to statements identified by roman numerals), lengthy word problems, and tables (which usually extend across two braille pages) may be more difficult to process because information takes longer to encode and must, therefore, be kept in short-

term memory longer. Another result is that units that can easily fit on a printed page must sometimes be broken up or reformatted. So, for example, questions must often be presented on a different page from the stimulus to which they refer and axis labels in graphs often are replaced with letters defined in a legend below the graph. Yet a third result is that figure labels (e.g., angle measurements) cannot be unambiguously placed unless the figure is substantially larger than the printed model.

A second effect of reading in braille is that some symbols that have no special meaning in print have meaning in braille and, thus, can cause confusion. In addition, some meaningful symbols that are not easily confused in print are so in braille. For example, the letters "A-J" and the numbers "1-9" use--with the exception of a prefix--the same braille notation.

For Black students, past research also offers hints about factors that might cause differential item difficulty. For example, Shepard, Camilli, and Williams (1984), using data from 10th and 12th graders participating in the High School and Beyond study, found verbally-loaded math items to show frequent indications of bias. During development of the Otis-Lennon Mental Ability Test, Scheuneman (1978) found math word problems to be more frequently biased on the 11th-12th grade form of the test than items involving straightforward number manipulation. Finally, Scheuneman (1985) discovered GRE Quantitative items requiring the problem to be abstracted from

a verbal description to show similar indications of differential difficulty. In addition to more abstract, verbally-loaded math problems, Scheuneman's (1985) study discovered indications of differential functioning associated with geometry items, key placement (Black students appeared to use a guessing strategy based on selecting the middle option), and quantitative comparison items that could be solved with a diagram but which did not include one.

Forming item categories. On the basis of the previous analysis, 13 overlapping Mathematical cluster structures were formed (structures were overlapping in that items belonged to more than one structure simultaneously). Within each structure, items were organized into mutually-exclusive content categories. These categories included ones hypothesized to be differentially difficult and "baseline" categories not hypothesized to show differential functioning (several categories, such as Diagram size: medium, were borderline ones meant to separate categories that clearly should show differential functioning from those that clearly should not). Table 3 presents the item categories composing each cluster structure and identifies those hypothesized to be differentially difficult. Definitions for each category are given in the appendix.

Insert Table 3 about here

Several points about Table 3 should be noted. First, a cluster structure was tested for a population only if an expectation of differential functioning existed for that group. Second, for some cluster structures, only a small number of item categories were hypothesized to function differentially. The remaining categories in the structure were tested to provide the baseline information referred to above. Finally, no research could be located to generate hypotheses about the specific types of geometry items that might function differentially for Black students. Consequently, all categories were tested as an exploratory endeavor.

For each of the categories listed in Table 3, one or more item clusters per form was constructed depending upon the number of available items meeting the category definition. In forming clusters, efforts were made to keep the number of items within the five-to-nine range. The five-item lower bound reduces the influence of guessing on item performance, making for more reliable behavioral indicators than could otherwise be obtained. In addition, with a minimum of six score points, a reasonable interval scale can be achieved. The maximum of nine items was set to keep individual clusters from becoming too large with respect to the total number of items in the test.

Because some theoretically-meaningful clusters had very few items (e.g., 3-dimensional solids), it was not always possible to maintain the five-item lower bound. Rather than

combine these instances into larger heterogeneous groupings, these "mini-clusters" were retained to be analyzed as clusters when composed of two-to-four items and as individual items when composed of a single test question.

Identifying differentially functioning clusters. To determine if item clusters operated differently for the study groups, a covariance adjustment was used in which the linear regression of each item-type cluster score on the total Mathematical score for each reference group was computed. In performing these computations, rights-only raw scores were used and the cluster score was removed from the total score. A comparison of the standardized difference between the study and reference group means on rights-only, formula, and scale scores suggested that these scores were functionally equivalent for group matching purposes (see Table 4).

Insert Table 4 about here

Using the reference-group regressions, cluster scores for each study group were predicted from their members' total Mathematical scores (after removing the cluster score). The predicted cluster mean for each group was then subtracted from that group's actual cluster mean, yielding a positive residual if the study group students did better than expected and a negative one when performance was lower than predicted. Finally, this residual was divided by the cluster standard deviation for the study group. A hypothesized cluster

category was said to be differentially difficult if its standardized residuals were equal to or less than -.2 on the majority of instances of that category across SAT forms and the associated baseline category generally showed no consistent evidence of differential difficulty. This criterion is suggested by Cohen, 1969, as a minimum for identifying the presence of meaningful effects in the social sciences. It is recognized that this criterion is somewhat arbitrary and that there is considerable debate over what size effect should be considered meaningful. However, previous analyses using this criterion have shown it to be a reasonably liberal one in identifying item effects (Bennett, Rock, & Kaplan, 1987).

In a few cases, clusters were composed of only single items. In these instances, the differential difficulty criterion was set at the approximate equivalent of a 10 percentage-point difference in the probability of passing the item (the statistic used to evaluate the functioning of individual items and the rationale for this criterion are described in the next section).

As noted, the procedure used is a form of covariance adjustment. In general, such adjustments require that assumptions of linearity and parallelism of regression be met. In the present case, the use of item cluster scores as the dependent variable decreases the possibility of nonlinearity because such scores are continuous. Additionally, where nonlinearity exists, linear regression estimates should

provide reasonable approximations. Further protection against nonlinearity in the reference sample might have been provided by matching reference and study group subjects on total score before estimating cluster performance. This matching, however, would have reduced sample size and, consequently, the statistical power of the analysis. As a result, it was not implemented.

With respect to the assumption of parallelism, the covariance adjustment used here follows the so-called "Belson model" (Belson, 1956; Snedecor & Cochran, 1980), in which the regression estimates from a larger comparison group are used to predict effects for a smaller treatment group. Under this model, parallel regressions are not assumed. Also, for the present study, the fact that the reference and study group regressions may not always be parallel is of little consequence because primary interest is in cluster performance differences at that point in the total score distribution where most study group individuals fall--that is, at the group mean. The focus, in other words, is only on whether there is a discrepancy between the performance of study group individuals and reference students operating at the same level as the majority of study group examinees.

Determining differential item performance. The items composing a cluster were analyzed individually if (1) the category had been hypothesized to show differential difficulty, and (2) a majority of the clusters in the category behaved deviantly, and (3) associated baseline categories

showed no consistent evidence of such functioning. Items were analyzed individually to determine if the category itself defined a potentially biased item type or, alternatively, if only a few aberrant items accounted for the unusual cluster performance.

For each item composing an errant category, Mantel-Haenszel (M-H) statistics were generated (Holland & Thayer, 1986). The M-H procedure is a form of 2x2x$\underline{s}$ contingency table analysis with two groups (study and reference) each categorized by success or failure on an item and matched on $\underline{s}$ categories (the categories are typically $\underline{s}$ score levels of a test). The M-H statistic compares the odds of success for the two groups and can be expressed as:

$$\alpha_{MH} = \frac{\Sigma_s R_{bs} W_{fs} / N_s}{\Sigma_s R_{fz} W_{bs} / N_s}$$

where

R = the frequency of right responses,

W = the frequency of wrong responses, and

N = the frequency of responses in stratum $\underline{s}$.

A frequently-used transformation for α_{MH} is to the delta-scale (Holland & Thayer, 1986). This scale provides an effect-size estimate of differential item performance. The transformation is:

$$\Delta_{MH} = -2.35 \ln (\alpha_{MH}).$$

For the M-H procedure, the matching criterion employed was SAT Mathematical rights-only raw score. This use of total score as a matching variable is supported by research which suggests that SAT Mathematical scores are unidimensional and that they generally have the same meaning across handicapped and nonhandicapped groups (Dorans & Lawrence, 1987; Rock, Bennett, & Kaplan, 1987; Willingham, Ragosta, Bennett, Braun, Rock, & Powers, 1988).

For both the Black/White and visually handicapped/ nonhandicapped comparisons, examinees were partitioned into 61 levels based on SAT-Mathematical score. In addition, a correction for differential speededness given by Schmitt, Bleistein, and Scheuneman (1987) was used. This correction accounts for some subgroups differentially reaching items at the end of the test. The adjustment redefines the proportion correct at each score level from the total number of students getting the item correct divided by all students taking the test to the total correct divided by only those students who reached the item.

Table 5 presents the mean number of items not reached and the mean omitted for the study and reference groups. Black students reached significantly fewer items than White examinees on all three forms, while visually impaired students, by virtue of receiving extra time, completed significantly more items than their sighted counterparts on two of the three forms. Visually impaired students also omitted more items than reference group students. However,

for visually impaired students, the SAT is more of a power test. As such, it is more likely that visually impaired students omit, not because they are unsure of the answers and are rushing to complete the exam as might their nonhandicapped counterparts, but because they have thoroughly considered the items and do not know the answers. Therefore, attempting to correct for these differential omit rates does not seem justified.

Insert Table 5 about here

In the current study, an item was considered to show differential functioning if its Δ_{MH} was different from zero at the .05 level of significance and was of a practically important size. Because no conventional criterion for practical importance in the context of differential item operation exists, any specification must be a judgment. For the delta scale transformations of the M-H statistic a difference of one unit has been suggested as a meaningful difference (Wild, 1987b). Except for very hard or very easy items, such a difference is approximately equal to 10 percentage points in the probability of passing an item (Wild, 1987b).

Relating item and cluster performance. Large item effects do not necessarily help to explain cluster performance. For instance, contrary effects (e.g., a differentially easy item in a differentially difficult

cluster), though relatively rare, not only fail to support the cluster result but may actually dampen the effects of items consistent with cluster findings.

For the hypothesized clusters found to be differentially difficult, each item result was compared to the effect shown by the cluster. An item effect was considered helpful in explaining cluster performance if it was concordant with the differential difficulty of the parent. To determine whether item-level results suggested differential functioning for the broad class represented by a hypothesized cluster, the number of differentially functioning items was tabulated. An item category was said to evidence pervasive differential functioning if at least half of the items in a majority of clusters showed concordant differential effects.

Evaluating the Cluster-Based Method

To evaluate the cluster-based method, several analyses were performed. First, the number of significant item performances detected in the supported hypothesized categories was tabulated and compared to the number of unique items causing those performances. This analysis was intended to assess the effect of overlapping cluster structures on the interpretability of results: a high ratio of significant performances to unique items suggests that the same small core of items may be causing several categories in different cluster structures to operate deviantly.

Second, the number of unique significant items detected in the supported hypothesized categories was compared with the

number of unique significant items that would have been detected had all differentially functioning categories been analyzed at the item level. This analysis speaks to the effect of using theoretical predictions to guide the analysis. If the theoretical predictions are meaningful, deviant categories for which no theoretical predictions were made should contain few additional differentially functioning items.

Third, the number of unique significant items detected in the supported hypothesized categories was compared with the number of unique significant items that would have been detected by an analysis of all test items. The intention of this analysis was to estimate how comprehensive the cluster-based method was in detecting differential item functioning. To do this, M-H indices were computed for all items for which these statistics had not been previously generated; as before, items were considered to function differentially if their Δ_{MH} equalled or exceeded 1.0 and differed from zero at the .05 level of significance.

Fourth, the direction of differential functioning suggested by the cluster-based method was compared with that indicated by the item-level analysis. Based on the literature review, the cluster-based method predicted substantial differential difficulty for both study groups. If supported by the cluster-based empirical results, this finding should be capable of confirmation through item level analysis: all

other things being equal, on the whole test, more items should be differentially difficult than differentially easy.

Finally, the meaning of the standardized residual statistic was assessed by computing its product-moment correlation with the cluster mean Δ_{MH} value (i.e., the mean Δ_{MH} taken over all items in a cluster) and by reviewing the size of the mean Δ_{MH} values for the supported categories.

Results

Visually Impaired Students

Table 6 presents the standardized residuals for all tested cluster categories. Of those 14 cluster categories hypothesized to show differential difficulty, six had a majority of their residuals exceeding the -.2 criterion in conjunction with baseline categories that did not show similar differential functioning. These six categories were Geometry: triangles, Spatial factor: possible spatial component, Spatial factor: estimation helpful in eliminating options, Stimulus format: figures, Stimulus format: graphs and tables, and Diagram size: small (the borderline category, Diagram size: medium, was also differentially difficult). The residuals for the Geometry: triangles category must be considered only weakly supportive of the hypothesis, however; while strong indications of differential difficulty were clear for the WSA forms, the residual for the CSA form was just as strongly positive, contradicting any argument of consistent differential difficulty. Of the remaining five cluster categories, the most consistent evidence for differential

difficulty was found for Stimulus format: figures, for which six of six clusters exceeded the -.2 criterion, Diagram size: small, for which three of three clusters were differentially difficult, and Spatial factor: estimation helpful, for which two of two met the cut-off.

Insert Table 6 about here

Of the 14 categories hypothesized to be differentially difficult, eight did not show meaningful evidence of such functioning either because their cluster residuals were not consistently differentially difficult or because, while they were, so were the clusters belonging to their associated baseline categories. Among the former were Geometry: 3-D solids, Miscellaneous: newly defined operations, Spatial factor: no figure, but drawing helpful, Reading load: high, Graphic placement: separated from item, and Shading: shaded. Cluster structures in which both the hypothesized and baseline categories were differentially difficult included Embedded figures and Label A-J.

Finally, there were two instances in which cluster structures had some baseline categories that were differentially difficult and others that were not. Both of these instances represented collections of items that did not fit under any well-defined grouping. These were Geometry: other geometry and Miscellaneous: other miscellaneous.

Table 7 presents the proportion of differentially difficult items in each supported hypothesized category (the differentially functioning borderline category, Diagram size: medium is also included). As the table indicates, in several instances at least half of the items were differentially difficult for a majority of clusters: Geometry: triangles, Spatial factor: estimation helpful in eliminating options, Stimulus format: figures, and Diagram size: small/medium. Again, the Geometry: Triangles category met the cut-off because of the behavior of items on the WSA forms; items on the CSA form showed little indication of differential difficulty.

Insert Table 7 about here

An indication of the extent to which these results are affected by the presence of overlapping cluster structures can be gained from a comparison of the number of significant item performances to the number of unique significant items. Over the six supported hypothesized categories and one supported borderline category, 17 individual item performances were significant for WSA3, 22 for WSA5, and 19 for CSA5. These significant performances were due to a small core of items--eight on each of the WSA forms and seven on CSA5--that repeatedly appeared in different cluster structures.

One means of assessing the extent to which theoretical predictions were helpful in guiding the analysis is by

comparing the number of unique significant items found in the hypothesized categories to the number found in all differentially functioning categories. When all significant categories were considered regardless of hypothesis, only one more significant item was detected on each of the WSA forms and none on CSA5.

To determine how comprehensive the cluster-based method was in detecting differential item functioning, the proportion of unique significant items located by the method was computed as a function of all unique significant items on all forms of the test. The supported hypothesized categories accounted for 57% (8 of 14), 53% (8 of 15), and 64% (7 of 11) of significant items, respectively.

Fourth, the direction of differential functioning suggested by the cluster-based method was compared with the results of analyzing all items. The cluster-based method predicted numerous instances of differential difficulty and located substantial supporting evidence. When all items on all three forms were analyzed, the number of differentially difficult items was almost three times the number of differentially easy ones (see Table 8).

Insert Table 8 about here

Finally, to better understand the meaning of the standardized residual, (1) the mean Δ_{MH} value for each cluster was correlated with the cluster standardized residual and (2)

the level of cluster differential difficulty indicated by the standardized residual was compared for each cluster with the degree of difficulty suggested by the mean Δ_{MH}. For all analyzed clusters, the correlations were .85, .95, and .98 for WSA3, WSA5, and CSA5, respectively, suggesting substantial overlap in the phenomenon being measured. With respect to level of differential difficulty, table 9 lists the mean Δ_{MH} values for the supported hypothesized categories. As the table indicates, these values generally support the inferences made from the standardized residuals: most of the supported hypothesized categories had a majority of their clusters showing significant differential difficulty (i.e., mean values greater than -1.00). In addition, as with the other indices, the Geometry: triangles category evidenced contradictory results (differential difficulty for the WSA forms but differential easiness for the CSA form).

Insert Table 9 about here

Black Students

Table 10 presents the standardized residuals for all cluster categories tested for Black students. None of the six hypothesized categories had a majority of its residuals exceeding the -.2 criterion. Three nonhypothesized categories, however, met the criterion. These were Spatial factor: possible spatial component, Reading difficulty: easy, and Concrete/abstract: concrete.

Insert Table 10 about here

Because none of the hypothesized categories proved significant, no individual items were detected via the cluster-based method. When all significant categories were considered regardless of hypothesis, two unique significant items each were detected on CSA5 and CSA7, and three on GSA2 (these seven items accounted for all the differentially difficult items on all three forms). Finally, when Δ_{MH} values were computed for all items on all forms, the number of differentially difficult items was half the number of differentially easy ones (see Table 11), supporting the general absence of differential difficulty indicated by the cluster-based method.

Insert Table 11 about here

Correlations between the mean Δ_{MH} values and the standardized residuals for all analyzed clusters were .86, .86, and .84 for CSA5, CSA7, and GSA2 respectively, somewhat lower than for the visually impaired groups.

Discussion

The purpose of this study was to develop, try out, and evaluate a theory-based method of detecting the underlying causes of differential difficulty. Two population subgroups taking SAT-M--visually impaired students administered braille

test editions and Black students--were chosen, and the literature on their mathematical test performance searched so that hypotheses about the causes of differential performance could be posed.

For the visually impaired group, the methodology was moderately successful. The data supported the hypothesis of differential difficulty in several SAT-M cluster categories. Statistical analysis of individual items bolstered these cluster-level results, producing a reasonable set of characteristics that might make for differential difficulty: items in which figures were presented as part of the stimulus, which had small-to-medium sized diagrams, or in which estimation was helpful in eliminating options. Item-level analysis of all questions on all forms confirmed the existence of substantial differential difficulty for these students. At the cluster level, the standardized residuals were very highly correlated with the cluster mean Δ_{MH}, suggesting that the two indices were measuring very similar phenomena. The differences between the two might be owed to the speededness correction applied to Δ_{MH} and/or to the fact that with such small samples, Δ_{MH} might be somewhat less reliable (since it is an average of individual item values).

For Black examinees, the method was effective in that it generally agreed with the results of the item-level analysis: neither method detected any consistent evidence of differential difficulty. In addition, the standardized

residuals and cluster mean Δ_{MH} values were highly correlated, though not as highly as in the visually impaired group.

While the cluster-based method showed moderate success, considerable limitations were apparent. For instance, in both populations the majority of hypotheses went unsupported (none were supported for the Black examinee samples), while several baseline categories were.

Why did the method fail to find differential functioning where it allegedly should and find it where it allegedly shouldn't? For both groups, the research base on the cognitive processes associated with mathematics skill development was found to be sparse and the results offered largely unreplicated, making for weak theoretical propositions. While such a limited base provides an acceptable foundation for exploratory analyses, it greatly restricts the power of predictions and the meaning that can be ascribed to results.

Besides weak theory, a second possible confounding factor is the standardized residual criterion established for identifying differential cluster functioning. Among the Black samples, two nonhypothesized categories were significant because of the presence of one or two differentially difficult items per cluster. In these cases, the categories clearly did not represent an aberrant item type. To reduce the influence of single items on cluster functioning, the criterion might be made a less liberal. Simulation research might help in identifying the criterion scores that have the greatest

likelihood of identifying categories with pervasive differential functioning while producing an acceptable number of false positives and negatives.

An additional shortcoming evident in the results from the visually impaired samples is that even when hypotheses are supported they may not be the actual causes of differential functioning; other dimensions common to the items in a category might have caused the observed effect. For example, in the Embedded cluster structure both the hypothesized category (items with embedded figures) and its negation (items with nonembedded figures) showed strong, consistent evidence of differential difficulty. The same situation held for the Label A-J structure. Clearly, the hypothesized dimensions are not at the root of the differential difficulty observed here. In both cases, a plausible explanation is that the items composing the structures were subsets of the Stimulus format: figures structure, which showed pervasive differential difficulty at the item level.

As suggested, the problem of overlapping structures is a considerable one. When such structures are posed, the same small core of differentially functioning items can cause categories in several structures to function aberrantly, making unclear what dimension is actually causing the observed effect. This limitation is partially a function of a weak theory of differential functioning for the subgroup, but also a result of the quasi-experimental nature of the method:

without experimental control of item characteristics, strong tests of hypotheses cannot be conducted.

Finally, the method was able to locate in the visually handicapped group only about half the items found to be differentially difficult by the item-level method and, in the Black group, none of seven items. For the Black student group, the number of significant items detected by the item-level method is at about the chance level. For the visually impaired group, the number is larger, again pointing to the limited utility of the research base and the resulting hypotheses in fully accounting for differential difficulty.

How might the cluster-based method be applied most productively? As noted, one major improvement is for targeting analyses at a small number of carefully selected hypotheses derived from a reasonably strong research base. In the absence of such a base, analyses can be no more than exploratory and limited interpretability should be expected to result. Second, to the extent possible, overlapping cluster structures should be avoided. When overlapping structures are indicated, the structure that is the more theoretically supportable should be selected. Third, results might be supplemented with more powerful methodologies. In particular, experimental designs are needed to allow more definitive tests of the hypotheses generated through quasi-experimental cluster-based designs. Such studies allow a degree of control of item content that cannot be achieved when working with intact tests--the type of manipulation needed to make stronger

inferences about the causes of differential difficulty. Also of value might be protocol analyses where a small number of subgroup members are asked to solve aloud problems from suspect item classes. Results can help to confirm hypotheses supported through quasi-experimental research or, alternatively, help build the database needed to specify hypotheses more effectively.

In sum, under the proper conditions, the cluster-based method would seem a potential incremental improvement over post-hoc, item-level approaches to differential functioning. The approach would appear better if for no other reason than that it is more conservative: the focus is on determining if the data endorse one's propositions rather than on constructing explanations to support the data. It is equally clear, however, that the method has important limitations that can be avoided best by applications in which a relatively strong set of predictions can be derived from a sound research base.

References

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. Journal of Educational Measurement, 24, 41-55.

Belson, W. A. (1956). A technique for studying the effects of a television broadcast. Applied Statistics, 5, 195-202.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Dorans, N. J. (1982). Technical review of item fairness studies: 1975-1979 (RR-82-90). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach (RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Dorans, N. J., & Lawrence, I. M. (1987). The internal construct validity of the SAT. Princeton, NJ: Educational Testing Service.

Fox, J. (1984). Linear statistical models and related methods. New York: John Wiley & Sons.

Holland, P. (1985, October). On the study of differential item performance without IRT. Paper presented at the Military Testing Conference, San Diego.

Johnson, M. L. (1984). Blacks in mathematics: A status report. Journal for Research in Mathematics Education, 15, 145-153.

Jones, L. V. (1984). White-Black achievement differences: The narrowing gap. American Psychologist, 39, 1207-1213.

Jones, L. V., Burton, N. W., & Davenport, E. C. (1984). Monitoring the mathematics achievement of Black students. Journal for Research in Mathematics Education, 15, 154-164.

Kulick, E. (1984). Assessing unexpected differential item performance of Black candidates of SAT Form CSA6 and TSWE Form E33 (SR-84-80). Princeton, NJ: Educational Testing Service.

Matthews, W., Carpenter, T. P., Lindquist, M. M., & Silver, E. A. (1984). The Third National Assessment: Minorities and mathematics. Journal for Research in Mathematics Education, 15, 165-171.

Rock, D. A., Bennett, R. E., & Kaplan, B. A. (1987). The internal construct validity of a college admissions test across handicapped and nonhandicapped groups. Educational and Psychological Measurement, 47, 193-205.

Rogers, H. J., & Kulick, E. (1986, April). An investigation of unexpected differences in item performance between Blacks and Whites taking the SAT. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Scheuneman, J. (1978, March). Ethnic group bias in intelligence test items. Paper presented at the meeting of the American Educational Research Association, Toronto.

Scheuneman, J. (1985). Exploration of causes of bias in test items (GREB No. 81-21P). Princeton, NJ: Educational Testing Service.

Schmeiser, C. B. (1981, April). An application of experimental design to the study of item bias. Paper presented at the annual meeting of the American Education Research Association, Los Angeles, California.

Schmitt, A. P., Bleistein, C. A., & Scheuneman, J. D. (1987, April). Determinants of differential item functioning for Black examinees on Scholastic Aptitude Test analogy items. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Snedecor, G. W., & Cochran, W. G. (1980). Statistical methods. Ames, IA: The Iowa State University Press.

Warren, D. H. (1981). Visual impairments. In J. M Kauffman and D. P. Hallahan (Eds), Handbook of special education. Englewood Cliffs, NJ: Prentice-Hall.

Wild, C. (1987a, March 25). Personal communication.

Wild, C. (1987b, April). Information on differential item functioning (DIF) procedures (Provisional). Princeton, NJ: Educational Testing Service.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H. I. Rock, D. A., & Powers, D. E. (Eds). (1988). Testing handicapped people. Boston, MA: Allyn & Bacon.

Witkin, H. A., Birnbaum, J., Lomonaco, S., Lehr, S., & Herman, J. L. (1968). Cognitive patterning in congenitally totally blind children. Child Development, 39, 767-786.

Table 1

Background Data for Visually Impaired (VI)

and Nonhandicapped (NH) Examinees

Form	N	% of Males	Mean Age	SAT-M Mean(SD)	SAT-V Mean(SD)	Mean Math Grade	Mean # Years of Math
WSA3							
VI	91	47%	18.2	381(112)	421(124)	2.5	3.1
NH	1110	46%	17.1	498(118)	451(111)	---	---
WSA5							
VI	96	55%	17.9	442(140)	444(136)	2.9	3.0
NH	1398	47%	16.7	486(114)	444(113)	---	---
CSA5							
VI	74	50%	18.3	402(127)	422(127)	2.9	3.7
NH	5507	44%	17.3	471(112)	436(103)	3.0	3.6

Note. Math grades and number of years of math are self-reported data taken from the Student Descriptive Questionnaire. Math grade is the grade received in the most recently taken math course. Number of years of math is the number the student expects to complete by the end of high school. For math grades, the Ns for visually handicapped students were 37, 53, and 26 for WSA3, WSA5, and CSA5, respectively. For number of years of math, the Ns were 39, 48, and 28. Data on these variables were not available for nonhandicapped examinees taking WSA3 or WSA5.

Table 2

Background Data for Black and White Examinees

Form	N	% of Males	Mean Age	SAT-M Mean(SD)	SAT-V Mean(SD)	Mean Math Grade	Mean # Years of Math
CSA5							
Black	446	34%	17.3	363(94)	345(92)	2.7	3.6
White	4405	44%	17.3	482(106)	447(98)	3.0	3.6
CSA7							
Black	834	36%	17.4	359(89)	332(89)	2.6	3.3
White	4798	46%	17.4	466(107)	431(98)	2.8	3.5
GSA2							
Black	705	36%	17.5	366(86)	335(88)	2.5	3.6
White	3985	48%	17.6	465(112)	422(98)	2.8	4.0

Note. For the CSA forms, math grade is the grade received in the most recently taken math course; for GSA2, it is the average of grades received in all math courses. Math grades and number of years of math are self-reported data taken from the Student Descriptive Questionnaire. Number of years of math is the number the student expects to complete by the end of high school.

Table 3
Item Cluster Structures and Cluster Categories Hypothesized to be Differentially Difficult

Cluster Structure & Category	Visually Impaired Students	Black Students
Geometry		
triangles	x	x
polygons		x
3-D solids	x	x
other geometry		x
Miscellaneous		
number properties		
newly defined operations	x	
other miscellaneous		
Spatial factor		
no figure, but drawing helpful	x	x
possible spatial component	x	
estimation helpful in eliminating options	x	
ordinary geometry		
Reading difficulty		
difficult		x
medium		
easy		
Concrete/abstract		
concrete		
abstract		x
Stimulus format: Picture		
figures	x	
graphs and tables	x	
Reading load		
high	x	x
medium		
low		
Key		
key "c"		
not key "c"		x
Graphic placement		
separated from item	x	
not separated		
Shading		
shaded	x	
not shaded		
Diagram size		
small	x	
medium		
large		
Embedded Figures		
embedded	x	
not embedded		
Label A-J		
A-J	x	

Table 4

Standardized Mean Differences Between Groups on Mathematical Raw and Scale Scores

Contrast	Standardized Difference		
	Rights-Only	Formula Score	Scale Score
Visually Impaired/ Nonhandicapped			
WSA3	.97	.97	1.00
WSA5	.44	.39	.39
CSA5	.65	.66	.62
Black/White			
CSA5	1.10	1.12	1.12
CSA7	.99	.99	.99
GSA2	.88	.91	.89

Note. Standardized differences were computed using the standard deviation of the appropriate reference group.

Table 5

Mean Number of Items Omitted and Not Reached for Study Groups

Visually Impaired Group				
	Mean # Not Reached		Mean # Omitted	
Form	Visually Impaired	Nonhandi-capped	Visually Impaired	Nonhandi-capped
WSA3	1.7[a]	2.8	8.1[c]	4.4
WSA5	1.3	1.5	8.2[d]	4.8
CSA5	1.2[b]	2.6	7.1[e]	4.6

[a] $p < .05$, z (one-tailed) = -2.05
[b] $p < .01$, z (one-tailed) = -3.21
[c] $p < .001$, z (two-tailed) = 6.19
[d] $p < .001$, z (two-tailed) = 5.59
[e] $p < .001$, z (two-tailed) = 4.25

Black Group				
	Mean # Not Reached		Mean # Omitted	
Form	Black	White	Black	White
CSA5	3.2[a]	2.5	4.9	4.6
CSA7	2.5[b]	1.9	5.3[d]	4.4
GSA2	2.8[c]	2.1	6.0	5.6

[a] $p < .001$, z (one-tailed) = 4.05
[b] $p < .001$, z (one-tailed) = 4.87
[c] $p < .001$, z (one-tailed) = 5.08
[d] $p < .001$, z (two-tailed) = 4.09

Table 6

Standardized Residuals for Visually Impaired Students taking the Braille Edition of SAT-M

Cluster Category	WSA3	WSA5	CSA5
Geometry			
triangles	-.44	-.35	.34
polygons	.27	.03	-.76
3-D solids	---	.02	.14
other geometry	-.25	-.29	-.12
Miscellaneous			
number properties	-.41	.07	.15
newly defined operations	-.42	NSI	NSI
other miscellaneous	-.77	-.28	-.49
Spatial factor			
no figure, but drawing helpful	NSI	.04	---
possible spatial component	-.35	-.05,-.26	-.42
estimation helpful in eliminating options	---	-.37	-.66
ordinary geometry	-.47,-.02	-.16	-.12, .41
Stimulus format: Picture			
figures	-.43,-.49	-.41,-.33	-.26,-.47
graphs and tables	NSI	-.41	-.25
Reading load			
high	-.34, .41	-.08, .18	.15,-.08
medium	-.13,-.49	-.07,-.16	.14, .08
low	-.01, .20	-.18, .07	.18,-.07
Graphic placement			
separated from item	-.06	-.47	-.07,-.73,.09
not separated	-.71,-.13	-.35,-.26	---
Shading			
shaded	.17	-.42	---
not shaded	-.49,-.51	-.26	---
Diagram size			
small	-.41	-.26	-.20
medium	---	-.48	-.52
large	.10	-.36	-.07
Embedded Figures			
embedded	-.16	-.58	-.41
not embedded	-.64	-.27	-.27,-.32
Label A-J			
A-J	-.16	-.55	-.42
not A-J	-.53	-.34	-.31

Note. NSI = a single, nonsignificant item.

Table 7

Proportion of Differentially Difficult Items in Each Supported Hypothesized Cluster Category for Visually Impaired Students

Cluster Category	WSA3	WSA5	CSA5
Geometry			
triangles	3/5	2/4	1/5
Spatial factor			
possible spatial component	3/6	1/4,3/6	2/6
estimation helpful in eliminating options	---	2/4	4/6
Stimulus format: Picture			
figures	3/6,5/8	2/5,4/7	3/5,3/7
graphs and tables	0/1	2/3	1/6
Diagram size			
small	3/5	2/3	2/6
medium	---	4/6	3/7

Note. An item was considered to be differentially difficult if its Δ_{MH} equalled or exceeded 1.0 and differed from zero at the .05 level of significance.

Table 8

Numbers of Differentially Difficult vs. Differentially Easy Items for Visually Impaired Students when all Items Across all Forms are Assessed

Form	Differentially Difficult	Differentially Easy
WSA3	14	3
WSA5	15	4
CSA5	11	8
TOTAL	40	15

Note. An item was considered to be differentially difficult if its Δ_{MH} equalled or exceeded 1.0 and differed from zero at the .05 level of significance. Differential easiness was defined as a Δ_{MH} equal to or less than -1.0 and differing from zero at the .05 level of significance.

Table 9

Mean Δ_{MH} Values of Supported Hypothesized Cluster Categories for Visually Impaired Students

Cluster Category	WSA3	WSA5	CSA5
Geometry			
triangles	-1.03	-1.24	.78
Spatial factor			
possible spatial component	-1.11	-.02,-1.11	-1.13
estimation helpful in eliminating options	-----	-1.28	-1.87
Stimulus format: Picture			
figures	-.91,-1.67	-1.36,-1.14	-.89,-1.12
graphs and tables	-.32	-2.02	-.52
Diagram size			
small	-1.65	-1.14	-.54
medium	-----	-1.60	-1.29

Table 10

Standardized Residuals for Black Students taking SAT-M

Cluster Category	CSA5	CSA7	GSA2
Geometry			
triangles	-.02	-.16	-.04
polygons	-.34	-.01	.24
3-D solids	.23	-.29	NSI
other geometry	-.08	.03	-.20
Spatial factor			
no figure, but drawing helpful	----	-.06	-.09,.09
possible spatial component	-.47	-.44	-.32
estimation helpful in eliminating options	-.33	-.03	NSI
ordinary geometry	.06	.09	.15
Reading difficulty			
difficult	-.16	-.14	.13
medium	-.18	-.26	.03
easy	-.28	.13	-.20
Concrete/abstract			
concrete	-.32,-.27	-.25,-.12	-.20,-.15
abstract	-.25,-.18	-.23,-.26	-.21,-.29
	-.25, .03	.23,-.26	-.10, .00
	.27, .44	.32, .07	.20, .24
Reading load			
high	-.07,-.14	-.07,-.10	.01,-.22
medium	-.08,-.03	-.14, .15	-.06, .05
low	-.07, .03	-.03,-.25	-.21, .06
Key (Regular math)			
key "c"	-.18, .26	-.16,-.17	.10,-.08
not key "c"	.02,-.21	-.29,-.10	-.10,-.06

Table 11

Numbers of Differentially Difficult vs. Differentially Easy Items for Black Students when all Items Across all Forms are Assessed

Form	Differentially Difficult	Differentially Easy
CSA5	2	4
CSA7	2	7
GSA2	3	3
TOTAL	7	14

Note. An item was considered to be differentially difficult if its Δ_{MH} equalled or exceeded 1.0 and differed from zero at the .05 level of significance. Differential easiness was defined as a Δ_{MH} equal to or less than -1.0 and differing from zero at the .05 level of significance.

Appendix: Cluster Category Definitions

Geometry

Triangles: item involves one or more triangles.

Polygons: item involves one or more polygons, other than triangles.

3-dimensional solids: item involves one or more 3-dimensional solids.

Other Geometry: item involves points, rays, lines, or angles in a plane; circles; or coordinate geometry (i.e., number line or rectangular coordinate system).

Miscellaneous

Number Properties: item concerns the structure of the number system or elementary number system concepts.

Newly defined operations: item contains special symbols or made up definitions.

Other miscellaneous: item concerns new concepts, probability, geometric perception, or sets.

Spatial Factor

No figure, but drawing helpful: item does not have a figure associated with it but making a sketch or drawing would help in solving it.

Possible spatial factor: item may require spatial skills to generate a solution.

Estimation helpful: spatial estimation appears helpful in eliminating at least two of the options.

Ordinary geometry: item can be solved by reference to factual relationships, rather than by spatial intuition.

Reading Difficulty (stem only)

Difficult: items containing compound sentences and/or large numbers of words perhaps requiring logic to sort out the meaning. Items which require careful reading.

> Example: Worker W produces n units in 5 hours. Workers V and W, working independently but at the same time, produce n units in 2 hours. How long would it take V alone to produce n units?

Medium: items with less verbiage; contain a simple word, phrase, or short sentences. Meaning is readily clear.

Example: A certain photocopying machine can make 10 copies every 4 seconds. At this rate, how many copies can the machine make in 6 minutes?

Easy: items which do not contain words or items which contain only a few (at most) standard words, such as (a) if ____, then _____ (b) _____ and _____, (c) if _____, and ____, then and (d) in the figure above.

Example: (a) If $y/x = -1$, then $y + x =$
(b) $x = 9$ and $y = 3$
(c) If $2x + 3y = 15$, and $y = 1$, then $2x =$
(d) In the figure above, $x =$
(without any further explanation given, other than the figure. If a more detailed explanation is given in the stem, the item would be considered to fall in the medium category.)

Concrete/Abstract

Concrete: questions which are real-life word problems.

Example: A supervisor was paid for her travel expenses at the rate of $0.20 per mile. If she received $14.40, for how many miles was she paid?

Abstract: questions that do not involve real-life settings.

Example: What is the sum of the areas of two squares with sides of lengths 1 and 3, respectively?

Stimulus Format: Pictures

Figures: item contains a figure or picture that does not have a coordinate system (has a triangle, square, rectangle, line segment, etc.).

Graphs and Tables: item has a coordinate system or is a line, bar or circle graph; is a number line; or item has data presented in rows and columns. The latter includes magic squares and times tables.

Reading Load

High: the number of words in the item is in the highest quartile for all items in the test section.

Medium: the number of words in the item is in the middle 50% of all items in the test section.

Low: the number of words in the item is in the lowest quartile for all items in the test section.

Key

Key "c": item is a five choice item with the correct answer corresponding to "c."

Not key "c": item is a five choice item with the correct answer corresponding to "a," "b," "d," or "e."

Graphic Placement

Separated from item: the item contains a figure that is placed on a page separate from the text of the item.

Not Separated: the item contains a figure that appears on the same page as the item proper.

Shading

Shaded: the item contains a shaded figure.

Not Shaded: the item contains a figure that is not shaded.

Figure Size

Small: the figure is among the smallest third in area across all figures across forms (less than 15 square inches)

Medium: the figure is among the middle third in area across all figures across forms (from 15 to 26 square inches)

Large: the figure is among the largest third in area across all figures across forms (more than 26 square inches)

Embedded Figures

Embedded: the item contains a geometric figure that has another geometric figure embedded in it to which the item refers.

Not Embedded: the item contains a figure that does not have another geometric figure within it.